Narodowy Korpus Języka Polskiego – teoria i praktyka. Fakty, mity, potrzeby[56]

Anna ANDRZEJCZUK
Wydawnictwo Naukowe PWN
ul. Postępu 18
02-676 Warszawa
Instytut Języka Polskiego PAN
anna.andrzejczuk@pwn.com.pl

Abstrakt: Wydaje się, że wszyscy twórcy korpusów przywiązują wagę do tego, żeby były one reprezentatywne i zrównoważone. Zaczynają się jednak pojawiać głosy, iż pojęcia te są mało precyzyjne. W niniejszym referacie autorka zamierza wyjść od przemyśleń na temat tych dwóch pojęć, zastanowić się, czy powinny być one ważne dla twórców korpusu i czy mamy jakąkolwiek alternatywę. Alternatywą może być stworzenie korpusu, którego dobór tekstów nie będzie niczego reprezentował poza samym sobą, a mianowicie teksty powinno się dobrać nie na zasadzie „reprezentatywności", ale na podstawie ich „celowości". Żeby określić celowość, należy się przyjrzeć potrzebom osób z nich korzystających. Należy też pamiętać, że z korpusów korzystają przede wszystkim nieinformatycy, zatem warto by było utworzyć narzędzie, które będzie miało łatwy, przyjazny dla użytkownika interfejs.

Dobrym wzorem będą tryby instalacji programów windowsowych. Instalatory często oferują co najmniej dwie możliwości wyboru: wersję standardową, dla mniej zorientowanego w opcjach i poniekąd we własnych potrzebach użytkownika, jak i wersję niestandardową dla użytkowników o wysokim stopniu świadomości własnych potrzeb, umożliwiającą zainstalowanie tylko tych składników, które są rzeczywiście potrzebne.

The National Polish Language Corpus – Theory and Practice. Facts, Myths and Needs.

Abstract: It may seem that all corpora creators aim at establishing representative and balanced corpora. But some think that those two concepts are not precise. The author analyses whether those two concepts are important for corpora creators and whether there is any alternative. One alternative is the creation of a corpus which would represent only itself – texts would be selected not on the basis of their 'representative character' but their 'aim'. In order to determine the aim of the text, it is necessary to investigate the needs of text users. It should be also born in mind that corpora users are usually not computer scientists and therefore they need a user friendly interface.

The installation mode of Windows may be a good example here. The installing software usually offers two installation modes: (i) a standard one for users who are not aware of their needs and (ii) an advanced one for users who know exactly what they need.

Teoria i praktyka

Początek niniejszego referatu jest daleki od oryginalności. Wpisuje się jednak w tradycję tekstów językoznawczych dotyczących tematyki korpusowej, a szczególnie tekstów poświęconych budowie korpusu.

[56] Praca naukowa finansowana ze środków na naukę w latach 2007–2010 jako projekt rozwojowy.

Zatem od czego zaczniemy? Jak przystało na naukowy wykład, zaczniemy od ustalenia terminologii, pojęć, jakimi zwykle posługują się twórcy korpusów i od jakich rozpocznę swoje właściwe wystąpienie. Te pojęcia to: *reprezentatywność* i *zrównoważenie.* Wydaje się oczywiste, że słowa te to dwa różne terminy. Zatem budzi zdziwienie fakt, że są one używane zamiennie i to nie tylko w języku polskim. Podobnie rzecz się ma z odpowiednikami angielskimi *representative* i *balanced*. Spróbujmy jednak po pierwsze – dokładnie sformułować definicje obu terminów, po drugie – odpowiedzieć na pytanie, dlaczego one są intuicyjnie stosowane zamiennie.

Co to znaczy, że korpus jest reprezentatywny? Czego korpus jest reprezentantem? Rafał Górski pisze, że korpus reprezentuje Saussurowskie parole – czyli rzeczywiste akty mowy, innymi słowy teksty już pomyślane, wypowiedziane i oczywiście – zapisane. W takim znaczeniu każdy korpus jest reprezentantem tekstów, tak jak reprezentantem ptaków jest każdy ptak, nawet jeśli nie jest typowym przedstawicielem gatunku. Nie każdy korpus jednak będzie reprezentatywny dla całego zbioru tekstów. Bycie reprezentatywnym dla całości wymaga od cząstki reprezentatywnej posiadania najbardziej typowych cech całości.

Skoro chcemy, żeby korpus był reprezentatywny, to pożądane jest, żeby zawierał przekrój tekstów najbardziej typowych z każdej odmiany danego języka. Pisze o tym m.in. Górski (Górski, 2008, s. 117–120), i podobnie Anna Andrzejczuk i Maciej Czupryniak: „Niezależnie od niejasności terminologicznych intuicyjnie przyjmowanym założeniem jest, by korpus jak najlepiej odzwierciedlał język – zwykle w takich jego odmianach i w takich proporcjach, jakimi posługujemy się i z jakimi spotykamy się w codziennym życiu: rozmawiając swobodnie czy oficjalnie, czytając, pisząc, słuchając radia, oglądając telewizję czy w końcu korzystając z Internetu." (Andrzejczuk, Czupryniak, 2009, s. 191).

Górski skupia się na tekstach wytwarzanych przez ludzi posługujących się danym językiem, a Andrzejczuk i Czupryniak wzięli pod uwagę teksty, które społeczeństwo nie tylko wytwarza, ale również odbiera. Oczywiste jest, że więcej tekstów jest przez ludzi odbieranych niż produkowanych, zatem te dwa podejścia mogą prowadzić do różnych zbiorów tekstów.

Drugim pojęciem, o którym wspomnieliśmy na początku, jest zrównoważenie. Zrównoważeniem nazwiemy dobór tekstów w takich proporcjach, które pozwolą nam uzyskać optymalną reprezentatywność. Zatem równoważenie jest procesem tworzenia korpusu, a zrównoważenie ukończeniem tego procesu.

Zwróćmy teraz uwagę, że korpus zrównoważony jest jednocześnie korpusem reprezentatywnym, ponieważ równoważymy do proporcji, jakie wymuszane są przez (bardzo różnie rozumianą) reprezentatywność. Im bardziej korpus jest reprezentatywny, tym lepiej jest zrównoważony. Nic więc dziwnego, że słowa *zrównoważony* i *reprezentatywny* stały się synonimami, ponieważ dotyczą tego samego stanu korpusu. Nie są to jednak synonimy dokładne, ponieważ zwracają uwagą na różne aspekty uzyskania ostatecznych wyników. *Zrównoważony* koncentruje się na procesie doboru, a *reprezentatywny* na sposobie ustalenia proporcji tekstów.

Dlaczego dążymy do reprezentatywności korpusów? W praktyce dzięki reprezentatywności chcemy się dowiedzieć, np.: jaki faktycznie jest język ogółu społeczeństwa nim się posługującego. Stworzenie takiej bazy umożliwia nam oparcie

analizy języka na danych obiektywnych (bo już użytych w aktach komunikacji językowej i zapisanych, a więc utrwalonych), a nie na subiektywnym, często chwilowym, zmiennym wyczuciu, jak się powinno mówić. Chcemy się dowiedzieć, jakie są typowe połączenia wyrazów, jakie słowa w polszczyźnie należą do najczęstszych, a jakie do najrzadszych, czy frekwencja słów odzwierciedla obraz rzeczywistości danej społeczności, czy w polszczyźnie wyraz *marchewka* jest zdecydowanie częstszy niż *avocado*, *ziemniaki* występują częściej niż *ryż*, *lekarz* częściej niż *marszałek*, *herbata* częściej niż *cola*, *rajstopy* częściej niż *pończochy*, bo Polacy częściej jedzą marchewkę i ziemniaki niż avocado i ryż, czy też czasami nad rzeczywistością frekwencyjnie biorą górę przyzwyczajenia językowe nad życiowymi, jak w przypadku *kawy* i *herbaty*[57].

Tutaj oczywiście nasuwają się kolejne pytania: w jaki sposób osiągnąć reprezentatywność? I czy faktycznie dzięki reprezentatywności uzyskujemy korpusy najlepiej przystosowane do naszych potrzeb? A może to nie jest najlepsze rozwiązanie? Czy istnieją inne? Jeżeli tak, to jakie?

Twórcy korpusów próbują osiągnąć reprezentatywność na różne sposoby, często przyjmując arbitralne założenia, uzasadniając, że dzięki takim a nie innym kryteriom, zbiór będzie mniej lub bardziej reprezentatywny i zrównoważony. Górski podaje kilka sposobów tworzenia korpusów (Górski, 2008, s. 117–118):

a) zbiór przypadkowych, a raczej przypadkowo osiągalnych tekstów,
b) zbiór równych ilości tekstów w wyznaczonych z góry kategoriach,
c) zbiór odzwierciedlający strukturę tekstów drukowanych,
d) zbiór odzwierciedlający produkcję tekstów,
e) zbiór odzwierciedlający strukturę czytelnictwa.

Największym korpusem typu (a) jest oczywiście Internet. Żaden zespół nie jest w stanie stworzyć większego korpusu. Przy takim ogromie tekstów tworzonych przez ludzi spontanicznie, korpus równoważy się samoistnie. Trudno też nie zgodzić się ze stwierdzeniem, że – choć Internet ma wiele wad – sztucznie tworzony korpus z takimi założeniami, będzie się miał do Internetu, jak gospodarka centralnie planowana do wolnorynkowej (Andrzejczuk, Czupryniak, 2009, s. 192).

Nietrudno zaś zauważyć plusy korpusu typu (b). Pawłowski udowadnia, że aby zbiór był reprezentatywny, musi być homogeniczny, czyli spełniać warunek jednorodności cech kwalifikacyjnych. Chodzi m.in. o to, że różnice między stylem pisanym a mówionym są tak wielkie, że zmieszanie tych dwóch odmian nie da nam obrazu żadnego z nich (Pawłowski, 2003, s. 167–168). Przy takim typie zrównoważenia tekstów plusem więc jest, np.: wyłapanie najczęstszego słownictwa, form oraz konstrukcji typowych dla danej kategorii. Minusem zaś brak danych, na temat zależności między nimi; danych, która kategoria pojawia się częściej, która rzadziej, która jest bardziej, a która mniej znana przeciętnemu Polakowi.

Wydaje się, że korpusy typu (c), (d) i (e) są próbą uzyskania struktury korpusu, o której mówiliśmy na początku. Chodziłoby o obiektywne kryteria, które miałyby przybliżyć nam świat słów, konstrukcji językowych, z którymi styka się, które zna

[57] Na łamach „Rzeczpospolitej" w *Słowach tygodnia* M. Łaziński uzasadniał, dlaczego w prasie częściej występuje *kawa* niż *herbata*. Przyczyna tego zjawiska tkwi nie w tym, że *kawa* jest częściej przez Polaków pita, tylko w tym, że *kawa* w polszczyźnie jest jednocześnie synonimem spotkania towarzyskiego.

przeciętny użytkownik posługujący się danym językiem. Próba utworzenia któregoś z tych korpusów wydaje się najbardziej sensowna. Zauważmy jednak, że to są trzy różne zbiory. Korpus (c) będzie podzbiorem korpusu (d). Natomiast teksty w korpusie typu (e) są dobierane z punktu widzenia odbiorcy, a nie nadawcy tekstu.

Pojawia się pytanie, co nam daje korpus reprezentujący stan czytelnictwa przeciętnych Polaków? Czy jest to pożądane w pracy leksykografa? Przypuśćmy, że książki czytaliby tylko uczniowie, a reszta społeczeństwa czytałaby jedynie prasę i ulotki reklamowe – czy wówczas korpus o właściwych proporcjach czytelnictwa byłby dobrany zgodnie z potrzebami leksykografa? Zanim przejdziemy do próby udzielenia odpowiedzi, przedstawmy najpierw obecnie tworzony największy korpus polszczyzny.

Fakty

Narodowy Korpus Języka Polskiego to projekt finansowany przez Ministerstwo Nauki i Szkolnictwa Wyższego. Tworzą go cztery instytucje, które już przed powstaniem projektu posiadały własne korpusy: Instytut Języka Polskiego PAN, Instytut Podstaw Informatyki PAN, Uniwersytet Łódzki i Wydawnictwo Naukowe PWN[58]. Połączenie zarówno doświadczeń, jak i zbiorów wszystkich partnerów pozwoliło na stosunkowo szybkie opublikowanie pierwszej wersji demonstracyjnej NKJP[59]. Obecnie opublikowana jest druga wersja demonstracyjna zawierająca 450 mln słów[60]. Do tej wersji weszły już teksty zbierane dla NKJP.

Docelowo cały korpus ma liczyć miliard słów. Siłą rzeczy, tak olbrzymi zbiór tekstów będzie oportunistyczny – będzie pełnił dla podkorpusu zrównoważonego rolę korpusu monitorowego, uzupełniającego. Ten największy korpus będzie miał jednocześnie budowę dwudzielną. Część tego korpusu będzie udostępniona do użytku publicznego na stronie www.nkjp.pl, część natomiast będzie tylko do użytku wewnętrznego, czyli w siedzibie wszystkich czterech instytucji tworzących NKJP[61]. Mniejszy podkorpus – rzędu wielkości trzystu milionów słów – ma być korpusem typu (e), a zatem jego zrównoważenie ma odzwierciedlać strukturę czytelnictwa w Polsce. Planowaną strukturę NKJP przedstawiono na rysunku 1.

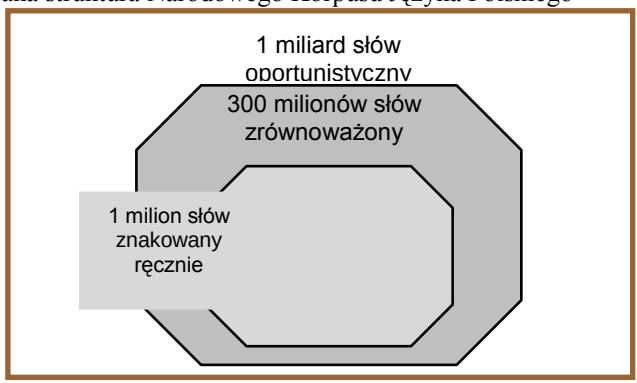
[58] Więcej informacji na stronie www.nkjp.pl

[59] Pojawiła się ona już w pół roku po rozpoczęciu projektu i zawierała ok. 350 mln słów.

[60] Dla porównania podaję liczbę słów w innych Korpusach Narodowych: Cambridge International Corpus – ok. 1,2 mld słów, British National Corpus – 100 mln słów, Český národní korpus – 300 mln słów, Nacional'nyj Korpus Russkogo Jazyka – 150 mln słów.

[61] Takie rozwiązanie podyktowane jest trudnością zdobycia praw do publicznego udostępniania choćby fragmentów niektórych tekstów.

Rys. 1. Planowana struktura Narodowego Korpusu Języka Polskiego



Jak już wspomnieliśmy, NKJP będzie zrównoważony według struktury czytelnictwa. Zatem w NKJP najwięcej będzie tekstów publicystycznych i wiadomości prasowych. Będą one stanowić aż 50% 300 mln korpusu. Pozostałe 50% to: literatura piękna (proza, poezja, dramat) – ok. 16%, literatura faktu – 5,5%, teksty informacyjno-poradnikowe – 5,5%, prace naukowo-dydaktyczne – 2%, teksty niebeletrystyczne, niesklasyfikowane – 1%, inne teksty pisane (m.in. ze stylu urzędowo-kancelaryjnego) – 3%, ponadto dodany zostanie znikomy procent tekstów ulotnych (ogłoszenia, reklamy, teksty propagandy politycznej, krótki instrukcje, listy). Nie należy również zaniedbywać coraz bardziej popularnych tekstów internetowych. Wejdzie ich do korpusu zrównoważonego ok. 7%. Będą to zarówno teksty spontanicznie tworzone przez użytkowników, takie jak fora, czaty, listy dyskusyjne, jak i teksty przemyślane, starannie zaplanowane, czyli teksty ze statycznych stron www. Ponadto, w NKJP ma być ok. 10% zapisanych tekstów mówionych. Większość będzie zapisem rozmów medialnych i quasi-mówionych[62]. Będzie też niewielki procent tekstów konwersacyjnych[63].

Mity

Jak już wspomnieliśmy, językoznawcy przywiązują do reprezentatywności i zrównoważenia korpusu bardzo dużą wagę. Twórcy korpusu zaś próbują się dostosować, wymyślając i opracowując różne schematy, mające prowadzić do spełnienia teoretycznego wymogu. Przypomina to trochę prześciganie się autorów słowników ortograficznych w umieszczeniu jak największej liczby haseł. Wydaje się, że w owych słownikach ważniejsza jest liczba słów niż faktyczny stopień ich trudności. Odpowiedź na pytanie, czy w słowniku ortograficznym potrzebne są słowa, z którymi przeciętny człowiek nie ma problemu, zostaje, chyba głównie ze względów marketingowych, przemilczana.

---

[62] Są to specjalnie zredagowane zapisy rozmów (jak w wywiadach prasowych) oraz teksty napisane w celu przemówienia.

[63] Po referacie, w kuluarach kilka osób pytało mnie, dlaczego w NKJP jest tak znikomy procent tekstów mówionych. Wiąże się to głównie z tym, że takie teksty są czasochłonne, wymagają sporych nakładów finansowych i pracy wielu osób. Ponadto, problem też jest zdobycie zgód na publiczne wykorzystywanie nagrań, pomimo, że są z nich usuwane wszelkie fragmenty identyfikujące rozmówcę.

Nie jestem pewna, czy owa reprezentatywność nie jest tylko pięknym mitem. Nie jestem w tej opinii osamotniona. Adam Pawłowski (2003, s. 20) dowodzi, że nie jest możliwy korpus reprezentujący zbiory otwarte, a do takich należy leksyka. Adam Kilgarriff i Gregory Grefenstette (2003, s. 343) stwierdzają wprost, że korpusy nie reprezentują niczego więcej poza sobą.

Wróćmy jednak teraz do wcześniej zadanego pytania. Co daje nam korpus reprezentujący strukturę czytelnictwa?

Niewątpliwie daje nam obraz, z jakim typem tekstów przeciętni Polacy są najbardziej oswojeni. Wiemy, jaki styl, jakie konstrukcje językowe, jakie słowa są dla nich najbardziej typowe, a z jakimi stykają się sporadycznie bądź w ogóle. Wiemy też, jakim językiem posługują się ludzie najbardziej do tego predysponowani, czyli pisarze, dziennikarze, naukowcy. W dzisiejszych czasach lingwista nie musi się już kierować jedynie własnym, subiektywnym wyczuciem przy opisie języka. Może teraz tworzyć opisy bardziej obiektywne, ponieważ ma możliwość sprawdzenia, jak używa języka szerszy przekrój społeczeństwa[64].

Potrzeby

Pytań jednak rodzi się więcej. Czy jakakolwiek reprezentatywność korpusu spełni oczekiwania osób z niego korzystających? Czy w ogóle wiemy, jakie są oczekiwania użytkowników? Czy warto się zastanowić nad potrzebami użytkowników przy budowaniu struktury korpusu?

Zaczniemy od odpowiedzi na ostatnie pytanie. Bez wątpienia warto. Dobrze jest wiedzieć, do jakich badań, zadań i celów korpusy będą wykorzystywane.

Żeby znać oczekiwania użytkowników korpusu, wiedzieć, w jakim stopniu spełniane są ich potrzeby, musimy przede wszystkim wiedzieć, kto z korpusów korzysta. Aby poznać precyzyjniejsze odpowiedzi na postawione pytania, zapewne trzeba by było przeprowadzić ankiety. Ponieważ jednak w niniejszym artykule proponuję rozważyć inny niż dotychczas sposób tworzenia korpusów i nie staram się podać ostatecznych procedur, posłużę się – jedynie dla ilustracji problemu – danymi intuicyjnymi i zapewne cząstkowymi.

Dwie najważniejsze grupy odbiorców to

a) „specjaliści językowi":

– osoby próbujące sformułować zasady ogólne języka, czyli językoznawcy; leksykografowie; programiści zajmujący się tworzeniem programów przetwarzających teksty,

– osoby znające świetnie język, ale chcący tę znajomość pogłębiać (np. wniknąć w styl, którego samemu na co dzień się nie używa, dowiedzieć się, jakie są najczęstsze kolokacje interesującego nas słowa), czyli ludzie pióra;

b) ludzie uczący się władania językiem, czyli mówienia, pisania po polsku:

– uczniowie,

– cudzoziemcy.

[64] Daleka jestem od twierdzenia, że jesteśmy w stanie stworzyć w pełni reprezentatywny korpus, ale – jak podczas jednej z rozmów stwierdził prof. W. Gruszczyński – może nie o to chodzi, by złapać króliczka, ale by gonić go?

Intuicyjnie wydaje się, że inny zbiór tekstów potrzebny jest dla osób z punktu (a), a inny dla osób z punktu (b) (choć oczywiście częściowo zbiory te mogą na siebie nachodzić).

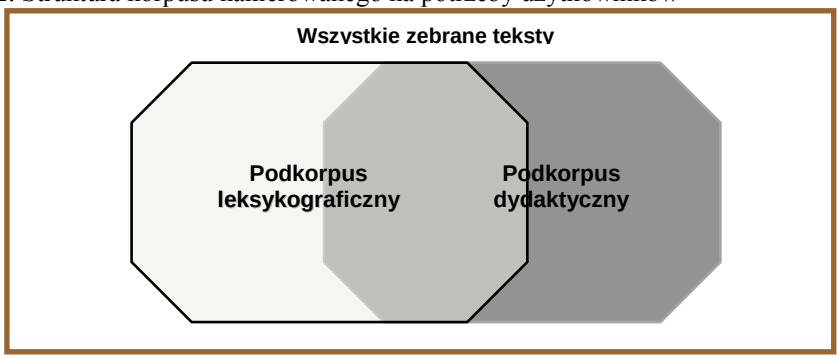
Osobom wymienionym w punkcie (a) potrzebny byłby korpus, który można by nazwać *leksykograficznym*. Oczywiście tego typu korpus powinny zaprojektować osoby mające doświadczenie zarówno w pracy z korpusami, jak i w leksykografii, czy w opisie zjawisk językowych. Nie wnikając głęboko w problem, można by pokusić się o stwierdzenie, że taki korpus powinien zawierać przekrój przez możliwie jak najwięcej odmian stylistycznych i tematycznych polszczyzny. W porównaniu z korpusem zrównoważonym powinien zawierać znacznie więcej książek, a mniej czasopism. Zdaje się, że w takim korpusie minimalną wartość mają zapisy tekstów mówionych przede wszystkim dlatego, że spontaniczne rozmowy cechuje zbyt duża skrótowość, nadmiernie wykorzystywane są nieokreślone zaimki wynikające z kontekstu niedostępnego dla osób trzecich. Nawet jak pojawiają się nowe wyrazy to, bez wiedzy na temat życia, przyjaciół, upodobań osób rozmawiających, często nie jesteśmy w stanie domyślić się znaczenia.

Natomiast osobom wymienionym w punkcie (b) potrzebny byłby korpus, który można by nazwać *dydaktycznym*. Korpus taki powinni zaprojektować metodycy, nauczyciele języka polskiego jako obcego oraz nauczyciele polskiego. Najprawdopodobniej, w odróżnieniu od korpusu leksykograficznego, ważną rolę pełniłyby tutaj zapisy nagrań mówionych (rozmów prywatnych, półoficjalnych, oficjalnych). Gdyby jeszcze technika na to pozwoliła i do tekstów dołączone byłyby ich zapisy dźwiękowe, to wówczas taki korpus, szczególnie dla osób dla których polski byłby językiem drugim, byłby bardzo cenny. W takim korpusie bardzo ważne byłoby otagowanie tekstów nie tylko stylistyczne, ale przede wszystkim tematyczne. W tym korpusie powinny być głównie teksty popularne, czasopisma codzienne i poświęcone różnym hobby. Tekstów naukowych tutaj można by było w ogóle nie włączać. W przypadku korpusu dydaktycznego bardzo ważna byłaby częsta aktualizacja.

Co jeszcze jest ważne przy tworzeniu publicznie dostępnych korpusów? Przede wszystkim należy pamiętać, że w przypadku tekstów, które chcemy wykorzystywać publicznie[65] wymagana jest zgoda właściciela praw autorskich. Dopiero 75 lat po śmierci autora teksty stają się własnością publiczną i niewymagana jest wówczas niczyja zgoda do ich wykorzystania. Nie tak łatwo jest pozyskać ogromną liczbę tekstów z pisemną zgodą od autorów, redaktorów i wydawców. Dlatego też dla twórców korpusu każdy tekst jest cenny i siłą rzeczy korpus musi być oportunistyczny. Jednakże zdanie się na całkowicie losowe proporcje dobieranych tekstów, rzecz jasna nie ma sensu. Dochodzimy zatem do wykluczających się założeń. Wbrew pozorom nie jest to jednak ślepa uliczka. Wystarczy stworzyć podkorpus z korpusu oportunistycznego. Na rysunku 2 przedstawiono strukturę korpusu ukierunkowanego na potrzeby użytkowników.

[65] Nie znaczy to wcale, że mamy zgodę na publikowanie w Internecie całych powieści, czy integralnych artykułów prasowych. Zgody są zwykle pisane tak, że w Internecie możemy pokazywać jedynie niewielkie fragmenty tekstów. Dzięki temu zarówno autorzy, jak i wydawcy nie muszą się bać, że czytelnicy zamiast kupić książkę czy gazetę, będą przeglądać korpus. Często może być wręcz przeciwnie. Użytkownik przypadkowo natknąwszy się na ciekawy fragment w korpusie, może pokusić się o dotarcie do źródła. Należy zauważyć, że wszystkie teksty w NKJP będą miały dokładne metryczki, które doprowadzą zainteresowanych do źródła tekstu.

Rys. 2. Struktura korpusu nakierowanego na potrzeby użytkowników



Oczywiście nie jesteśmy w stanie określić szczegółowych preferencji wszystkich użytkowników. Dlatego pod tym względem dobrym wzorem są tryby instalacji programów windowsowych. Zwróćmy uwagę, że instalatory często oferują co najmniej dwie możliwości wyboru: wersję standardową, dla mniej zorientowanych w opcjach i poniekąd we własnych potrzebach użytkowników, jak i wersję niestandardową dla użytkowników o wysokim stopniu świadomości własnych potrzeb, umożliwiającą zainstalowanie tylko tych składników, które są im rzeczywiście potrzebne. Nietrudno zgadnąć, że statystycznie częściej używana jest opcja standardowa, jednak takie rozwiązanie nie zaniedbuje potrzeb znacznie mniejszej, ale za to bardziej świadomej i co za tym idzie – lepiej wykorzystującej możliwości programu – grupy użytkowników.

Zatem, przy tak zwanych „gotowcach", czyli wersjach standardowych dla najbardziej charakterystycznych grup docelowych, powinna istnieć możliwość doboru tekstów w proporcji, którą zaproponuje sam zainteresowany. Jeżeli metadane dotyczące tekstu będą bardzo dobrze opisane, to wówczas nie powinno być większych problemów z napisaniem programu dobierającego teksty w proporcjach, których sobie będzie życzył użytkownik. W takich przypadkach wielkość korpusu byłaby dostosowana do wielkości dostępnych zasobów w żądanych proporcjach. Bardzo ważne też jest, żeby użytkownik mógł na własnym komputerze zapisać zaprojektowany przez siebie korpus, aby miał możliwość powrotu do niego.

Bibliografia

Andrzejczuk Anna, Czupryniak Maciej, 2008, *O wykorzystaniu zasobów internetowych w pracy językoznawcy*, [w:] „Polonica", t. 29, Kraków, s. 189–204.

Górski Rafał Ludwik, 2008, *Charakterystyka chronologiczna i stylistyczna korpusu dla „Wielkiego słownika języka polskiego"*, [w:] *Nowe studia leksykograficzne 2*, red. P. Żmigrodzki, R. Przybylska, Kraków, Wydawnictwo Lexis, s. 117–127.

Kilgarriff Adam, Grefenstette Gregory, 2003, *Introduction to the Special Issue on the Web as Corpus*, „Computational Linguistics", http://www.mitpressjournals.org/toc/coli/29/3.

Pawłowski Adam, 2003, *Uwagi na temat korpusu języka polskiego (reprezentatywność, aktualność, nazwa)*, [w:] *Językoznawstwo w Polsce. Stan i perspektywy*, red. S. Gajda, Opole, Polska Akademia Nauk – Komitet Językoznawstwa, Uniwersytet Opolski – Instytut Filologii Polskiej, s. 162–180.